(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－51890
（P2001－51890A）
(43)公開日　平成13年２月23日(2001.2.23)

(51)Int.Cl.[7]	識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ　12/00	５４５	Ｇ０６Ｆ　12/00	５４５Ａ	５Ｂ０８２
13/00	３５１	13/00	３５１Ｅ	５Ｂ０８９

審査請求　未請求　請求項の数７　ＯＬ　（全 14 頁）

(21)出願番号　特願平11－226494

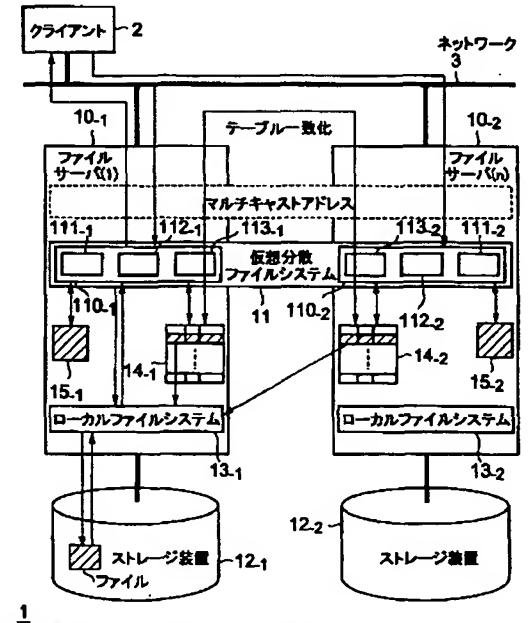
(22)出願日　平成11年８月10日(1999.8.10)

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　内堀　郁夫
東京都府中市東芝町１番地　株式会社東芝府中工場内
(72)発明者　高桑　正幸
東京都府中市東芝町１番地　株式会社東芝府中工場内
(74)代理人　100058479
弁理士　鈴江　武彦　（外６名）

最終頁に続く

(54)【発明の名称】　仮想分散ファイルサーバシステム

(57)【要約】
【課題】ネットワーク上に分散した複数のファイルサーバの台数やストレージ装置の接続状態をクライアントに意識させないで済むようにする。
【解決手段】ネットワーク３上に分散したサーバ１０-1，１０-2を備え、各サーバには、仮想分散ファイルシステム１１が分散して実装されている。このシステム１１を構成する、サーバ１０-1，１０-2上のモジュール１１０-1，１１０-2は、クライアント２からマルチキャストされたファイル操作要求を受け取ると、仮想分散ファイルシステム１１と全ローカルファイルシステム１３-1，１３-2とのマッピングテーブル１４-1，１４-2または全サーバのサーバ情報を保持するサーバ情報保持部１５-1，１５-2をもとに、自サーバが上記要求を処理可能な最適なサーバであるか否かを判断し、その判断結果に基づいて要求されたファイル操作を対応するサーバのローカルファイルシステムにより行わせる。



1
仮想分散ファイルサーバシステム

14-1,14-2…マッピングテーブル
15-1,15-2…サーバ情報保持部
110-1,110-2…仮想分散ファイルモジュール
111-1,111-2…仮想分散ファイルインタフェース
112-1,112-2…ローカルファイルインタフェース
113-1,113-2…通信モジュール

【特許請求の範囲】

【請求項１】　マルチキャスト可能なネットワーク上に分散した複数のファイルサーバを備えた仮想分散ファイルサーバシステムであって、

前記各ファイルサーバに分散して実装され、全ファイルサーバのファイルを統合的に管理する、実際のストレージ構成には非依存の仮想分散ファイルシステムと、

前記各ファイルサーバにそれぞれ独立して実装され、各サーバに固有のストレージ構成を管理するローカルファイルシステムと、

前記各ファイルサーバにそれぞれ設けられ、前記仮想分散ファイルシステムで統合的に管理される各ファイルについて、当該仮想分散ファイルサーバシステムとそのファイルを実際に管理する前記ローカルファイルシステムとの間のマッピングの情報を保持するマッピングテーブルとを具備し、

前記仮想分散ファイルシステムは、前記各ファイルサーバにそれぞれ設けられた管理モジュールから構成され、

前記各管理モジュールは、クライアントからマルチキャストされたファイル操作要求を共通に受け取り、当該要求に応じて自サーバの前記マッピングテーブルを参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断し、最適なサーバであると判断した場合だけ、要求されたファイル操作を対応するサーバの前記ローカルファイルシステムにより行わせるように構成されていることを特徴とする仮想分散ファイルサーバシステム。

【請求項２】　マルチキャスト可能なネットワーク上に分散した複数のファイルサーバを備えた仮想分散ファイルサーバシステムであって、

前記各ファイルサーバに分散して実装され、全ファイルサーバのファイルを統合的に管理する、実際のストレージ構成には非依存の仮想分散ファイルシステムと、

前記各ファイルサーバにそれぞれ独立して実装され、各サーバに固有のストレージ構成を管理するローカルファイルシステムと、

前記各ファイルサーバにそれぞれ設けられ、前記仮想分散ファイルシステムで統合的に管理される各ファイルについて、当該仮想分散ファイルサーバシステムとそのファイルを実際に管理する前記ローカルファイルシステムとの間のマッピングの情報を保持するマッピングテーブルと、

前記各ファイルサーバにそれぞれ設けられ、全ての前記ファイルサーバについて、そのサーバのストレージ装置の空き容量を示す情報、及びそのサーバの負荷状況を示す情報の少なくとも一方を含むサーバ情報を保持するサーバ情報保持手段とを具備し、

前記仮想分散ファイルシステムは、前記各ファイルサーバにそれぞれ設けられた管理モジュールから構成され、

前記各管理モジュールは、クライアントからマルチキャストされたファイル操作要求を共通に受け取り、当該要求に応じて自サーバの前記マッピングテーブルまたは前記サーバ情報保持手段を参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断し、最適なサーバであると判断した場合だけ、要求されたファイル操作を対応するサーバの前記ローカルファイルシステムにより行わせるように構成されていることを特徴とする仮想分散ファイルサーバシステム。

【請求項３】　前記管理モジュールは、前記ファイル操作要求がファイル読み出し要求またはファイル書き込み要求の場合には、自サーバの前記マッピングテーブルを参照し、該当するファイルが自サーバの前記ローカルファイルシステムの管理下にあるか否かにより、自サーバが前記要求を処理可能な最適なサーバであるか否かを判断することを特徴とする請求項１または請求項２記載の仮想分散ファイルサーバシステム。

【請求項４】　前記管理モジュールは、前記ファイル操作要求がファイルの新規作成要求の場合には、自サーバの前記サーバ情報保持手段を参照し、全ての前記サーバの各々について、そのサーバのストレージ装置の空き容量、またはそのサーバの負荷状況を比較することで、自サーバが前記要求を処理可能な最適なサーバであるか否かを判断することを特徴とする請求項２記載の仮想分散ファイルサーバシステム。

【請求項５】　前記管理モジュールは、全ての前記ファイルサーバの前記マッピングテーブルの内容を一致化するために、自サーバの前記マッピングテーブルの情報と他のサーバの前記マッピングテーブルの情報とをサーバ間通信により交換することを特徴とする請求項１記載の仮想分散ファイルサーバシステム。

【請求項６】　前記管理モジュールは、全ての前記ファイルサーバの前記マッピングテーブルの内容を一致化するために、自サーバの前記マッピングテーブルの情報と他のサーバの前記マッピングテーブルの情報とをサーバ間通信により交換する一方、全ての前記ファイルサーバの前記サーバ情報保持手段の内容を一致化するために、自サーバの前記サーバ情報保持手段の情報と他のサーバの前記サーバ情報保持手段の情報とをサーバ間通信により交換することを特徴とする請求項２記載の仮想分散ファイルサーバシステム。

【請求項７】　前記各ファイルサーバにそれぞれ設けられ、そのサーバの管理下にある各ファイル別の負荷状況を示す情報を保持するファイル別負荷状況情報保持手段を更に具備し、

前記管理モジュールは、自サーバの前記ファイル別負荷状況情報保持手段に保持されている情報から第１の閾値を超えた負荷のファイルを検出して、他の任意のファイルサーバに対してサーバ間通信により当該ファイルのレプリケーションを行い、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当

該要求に対する処理をレプリケーション側に任せるようにしたことを特徴とする請求項１または請求項２記載の仮想分散ファイルサーバシステム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、コンピュータ・ネットワークシステムにおけるファイルサーバシステムに係り、特にネットワーク上に接続された複数のファイルサーバを協調動作させて、クライアントからは単一のサーバとして機能させる仮想分散ファイルサーバシステムに関する。

【０００２】

【従来の技術】今日のコンピュータ・ネットワークシステムにおいては、ネットワークに接続された異なるコンピュータ間でファイルを共有することが一般的に行われている。こうした環境下では、特定のコンピュータに大規模なストレージを接続して、ファイルサーバとして運用したり、最近ではＮＡＳ（Network Attached Storage）と呼ばれる、ファイルサーバ専用機を接続する等のシステム構成をとることが多い。

【０００３】ファイルサーバを使用する環境（ファイルサーバシステム）では、サーバのストレージ容量が不足した場合には、サーバ側に物理的・性能的に拡張性があれば、新たにディスク装置等（のストレージ装置）を増設することで対処できる。このときクライアントからは、新たなボリュームをマウントして使用するといった形態になる。また、サーバの拡張性が限界に達していれば、サーバ自体を増設することになる。このときクライアントからは、増設したサーバを意識した上で新たなボリュームをマウントして使用するといった形態になる。

【０００４】

【発明が解決しようとする課題】上記したコンピュータ・ネットワークシステムにおいてファイルサーバを利用してファイル共有を行う場合、クライアント側からは、ファイルサーバ側のボリューム構成がそのまま見えてしまうのが一般的である。例えばサーバ側でディスク装置を増設した場合には、クライアント側は新たなボリュームを認識した上で、マウントしなければならない。或いはサーバ自体を増設した場合には、増設したサーバの運用ポリシーを決定、もしくはシステム設定・管理等の煩雑な作業が発生する上、クライアント側でも、新たなサーバを認識した上で、新たなボリュームをマウントしなければならない。

【０００５】このように従来のファイルサーバを用いたファイル共有システム（ファイルサーバシステム）では、ディスク装置（ストレージ装置）の増設、或いはサーバの増設が必要な場合、サーバ側、クライアント側のいずれにも、新たな設定・管理のために多大なコストが発生するという問題があった。更に、ストレージの利用形態よっては、特定のファイルシステムをそのまま容量だけ拡張したい場合もあり、単にストレージ装置やサーバを増設するだけでは解決しないケースもあった。

【０００６】本発明は上記事情を考慮してなされたものでその目的は、ネットワーク上に分散した複数のファイルサーバを、クライアントからは単一のサーバとして扱うことができ、サーバ台数やストレージ装置の接続状態をクライアントに意識させない仮想分散ファイルサーバシステムを提供することにある。

【０００７】

【課題を解決するための手段】本発明は、マルチキャスト可能なネットワークに接続された複数のファイルサーバに分散して実装され、全ファイルサーバのファイルを統合的に管理する、実際のストレージ構成には非依存の仮想分散ファイルシステムと、各ファイルサーバにそれぞれ独立して実装され、各サーバに固有のストレージ構成を管理するローカルファイルシステムと、前記各ファイルサーバにそれぞれ設けられ、上記各ファイルについて、仮想分散ファイルサーバシステムとそのファイルを実際に管理するローカルファイルシステムとの間のマッピングの情報（例えば、仮想分散ファイルサーバシステムで管理され、クライアントから見える仮想的なパスと、ローカルファイルシステムで管理され、クライアントから見えない物理的な所在とを対応付けた情報）を保持するマッピングテーブルとを備えると共に、上記仮想分散ファイルシステムを、各ファイルサーバにそれぞれ設けられた管理モジュールであって、クライアントからマルチキャストされたファイル操作要求を共通に受け取り、当該要求に応じて自サーバのマッピングテーブルを参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断し、最適なサーバであると判断した場合だけ、要求されたファイル操作を対応するサーバのローカルファイルシステムにより行わせる管理モジュールにより構成することを特徴とする。

【０００８】ここで、各ファイルサーバ上に、全ファイルサーバについて、そのサーバのストレージ装置の空き容量を示す情報、及びそのサーバの負荷状況を示す情報の少なくとも一方を含むサーバ情報を保持するサーバ情報保持手段を更に設け、上記各管理モジュールでは、クライアントからマルチキャストされたファイル操作要求を受け取った場合に、当該要求に応じて自サーバのマッピングテーブルまたはサーバ情報保持手段を参照することで、自サーバが当該要求を処理可能な最適なサーバであるか否かを判断する構成としてもよい。

【０００９】このような構成においては、クライアントから特定のファイルサーバを意識することなくマルチキャストされたファイル操作要求は、仮想分散ファイルサーバシステムを構成する各ファイルサーバ上の管理モジュールで共通に受け取られ、その要求に応じて対応するサーバ（自サーバ）のマッピングテーブルまたはサーバ情報保持手段が参照される。そして、この参照の結果、

自サーバが上記要求を処理可能な最適なサーバであるか否かが判断され、最適なサーバであると判断した唯一のサーバ（上の管理モジュール）だけが、要求されたファイル操作を自サーバのローカルファイルシステムにより行わせる。

【００１０】このように、要求元のクライアントからは、ネットワーク上に分散した複数のファイルサーバを単一のサーバとして扱うことができ、サーバ台数やストレージ装置の接続状態を意識する必要がない。

【００１１】ここで、上記管理モジュールで、自サーバが最適なサーバであるか否かを判断するためのアルゴリズムとして、以下の第１乃至第４のアルゴリズム（判断手法）のいずれかを適用するとよい。

【００１２】第１のアルゴリズムは、ファイル操作要求がファイル読み出し要求またはファイル書き込み要求の場合に適用されるもので、自サーバのマッピングテーブルの情報に基づいて、該当するファイルが自サーバのローカルファイルシステムの管理下にあるか否かにより判断する手法である。

【００１３】第２のアルゴリズムは、ファイル操作要求がファイルの新規作成要求の場合に適用されるもので、自サーバのサーバ情報保持手段の情報に基づいて、全てのサーバの各々について、そのサーバのストレージ装置の空き容量（空き記憶容量）、またはそのサーバの負荷状況を比較することで判断する（例えば、自サーバの空き容量が最も大きい場合、或いは自サーバの負荷が最も低い場合に上記最適サーバと判断する）手法である。

【００１４】第３のアルゴリズムも、ファイル操作要求がファイルの新規作成要求の場合に適用されるもので、自サーバのマッピングテーブルの情報に基づいて、全てのサーバの各々について対応するストレージ装置上に確保可能な連続領域を求め、その連続領域のサイズを比較することで判断する（例えば、自サーバのストレージ装置上に確保可能な連続領域のサイズが最も大きい場合に上記最適サーバと判断する）手法である。

【００１５】第４のアルゴリズムも、ファイル操作要求がファイルの新規作成要求の場合に適用されるもので、全てのサーバの各々について、そのサーバのストレージ装置の空き容量、そのサーバの負荷、及び当該ストレージ装置上に確保可能な連続領域の少なくとも２つを求め、その求めた少なくとも２つの情報を複合条件として比較することで判断する手法である。

【００１６】以上の第１乃至第４のアルゴリズムのいずれか１つを適用することで、クライアントから特定のファイルサーバを意識することなくマルチキャストされたファイル操作要求を各サーバが共通に受け取っても、その要求されたファイル操作を行うのに最適なサーバであるか否かを、その都度相互に通信を行うことなく、そのサーバ自身で自律的に判断することができる。

【００１７】ここで、上記各管理モジュールに、全てのファイルサーバのマッピングテーブルの内容を一致化するために、自サーバのマッピングテーブルの情報と他サーバのマッピングテーブルの情報とをサーバ間通信により交換する機能（通信モジュール）を持たせるとよい。また、マッピングテーブルに加えてサーバ情報保持手段を各サーバ上に備えた構成では、各管理モジュール（内の通信モジュール）に、全てのファイルサーバのサーバ情報保持手段の内容を一致化するために、自サーバのサーバ情報保持手段の情報と他のサーバのサーバ情報保持手段の情報とをサーバ間通信により交換する機能を更に持たせるとよい。

【００１８】また、マッピングテーブルの一致化のためには、自サーバのローカルファイルシステムで実際に管理されるファイル構成が変更された場合に、その変更された情報（マッピング情報）をサーバ間通信により他の全サーバに送信するのが効率的である。同様に、サーバ情報保持手段の内容の一致化のためには、自サーバのサーバ情報を定期的に更新し、その都度、その更新されたサーバ情報をサーバ間通信により他の全サーバに送信するのが効率的である。

【００１９】また本発明は、上記仮想分散ファイルサーバシステムにサーバが動的に増設された場合に、そのサーバの管理モジュールで以下の第１乃至第４の処理を行うようにしたことをも特徴とする。まず、第１の処理では、サーバ間通信により他の全てのサーバに対してマッピングテーブル及びサーバ情報保持手段の更新を禁止するロック設定を行い、次の第２の処理では、サーバ間通信により他の任意のサーバからマッピングテーブルサーバ情報保持手段の内容を自サーバにコピーし、次の第３の処理では、自サーバのサーバ情報保持手段に自サーバのサーバ情報を追加し、次の第４の処理では、サーバ間通信により自サーバのーバ情報を他の全てのサーバのサーバ情報保持手段に反映させて全サーバのサーバ情報保持手段の一致化を図り、しかる後に上記ロック設定を解除する。

【００２０】このようなサーバ増設時の一連の動作により、動的にサーバ台数を拡張できる。しかもクライアントは、サーバ台数の拡張を意識することなく、増設されたサーバを利用することができる。

【００２１】また本発明は、各ファイルサーバに、そのサーバの管理下にある各ファイル別の負荷状況を示す情報を保持するファイル別負荷状況情報保持手段を付加し、各サーバの管理モジュールにおいて、自サーバのファイル別負荷状況情報保持手段に保持されている情報から第１の閾値を超えた負荷のファイルを検出して、他の任意のファイルサーバに対してサーバ間通信により当該ファイルのレプリケーションを行い、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当該要求に対する処理をレプリケーション側に任せるようにしたことをも特徴とする。

【００２２】このような構成においては、自律的な負荷分散が可能となる。ここで、上記検出したファイルがレプリケーションされたファイルである場合にも、他の任意のファイルサーバに対してサーバ間通信により当該ファイルのレプリケーションを行い、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合に、当該要求に対する処理を新たなレプリケーション側に任せることにより、自律的な負荷分散がより広範囲に効果的に行える。また、他サーバへのレプリケーションの対象となったファイルの負荷が前記第１の閾値より低い第２の閾値以下となった場合に、そのレプリケーションを行ったサーバ上の管理モジュールから当該他サーバに対してサーバ間通信により対応するファイルの消去を要求し、クライアントからマルチキャストされた当該ファイルの読み出し要求があった場合、当該要求に対する処理を自身が行うことにより、動的な負荷分散が可能となる。

【００２３】さて、上記一致化のためのサーバ間通信（マッピング情報またはサーバ情報の通信）、更には上記レプリケーションのためのサーバ間通信には、上記ネットワークを用いることが可能である。しかし、各ファイルサーバを相互接続する専用の通信路（プライベート通信路）を上記ネットワークから独立に設け、当該通信路を用いてサーバ間通信を行う構成とするとよい。この場合、サーバ間通信のためにネットワークのスループットが悪化するのを防止できる。

【００２４】また本発明は、各ファイルサーバ及び当該サーバのストレージ装置を相互接続するマルチホストが可能なインタフェースを更に備えると共に、上記各管理モジュールによる上記サーバ間通信を当該インタフェースを介して行うようにしたことをも特徴とする。

【００２５】このような構成においては、上記各ストレージ装置を上記インタフェースによって各サーバ間で共有し、上記一致化のためのサーバ間通信、更には動的なファイルのレプリケーションのためのサーバ間通信が上記インタフェースを通して行われるため、自律的な負荷分散が効果的に実現される。

【００２６】なお、本発明は方法に係る発明としても成立する。

【００２７】

【発明の実施の形態】以下、本発明の実施の形態につき図面を参照して説明する。

【００２８】［第１の実施形態］図１は本発明の第１の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図である。

【００２９】同図において、１は仮想分散ファイルサーバシステム、２は同ファイルサーバシステム１にファイルサービスを要求するクライアント（クライアントコンピュータ）である。仮想分散ファイルサーバシステム１は、ネットワーク３上に分散配置された複数、例えば２台のファイルサーバ（サーバコンピュータ）１０-1，１０-2を用いて実現される。なお、図ではクライアント２は便宜的に１台だけが示されているが、複数存在するのが一般的である。

【００３０】１１は仮想分散ファイルサーバシステム１の中心をなす仮想分散ファイルシステムであり，各ファイルサーバ１０-1，１０-2に分散して実装されている。この仮想分散ファイルシステム１１は、全ファイルサーバ１０-1，１０-2のファイルを統合的に管理し、ファイルサーバ１０-1，１０-2それぞれの実際のボリューム構成（ストレージ構成）には依存しない、仮想的なファイルシステムをクライアント２に対して提供するものである。

【００３１】仮想分散ファイルシステム１１は、各ファイルサーバ１０-1，１０-2に分散して実装された仮想分散ファイルモジュール１１０-1，１１０-2を有している。仮想分散ファイルモジュール１１０-1，１１０-2は、ファイルサーバ１０-1，１０-2上でクライアント２からの要求を分散して処理しつつ、クライアント２に対しては仮想的に１つのファイルシステムとして見せるための管理モジュールである。仮想分散ファイルモジュール１１０-1，１１０-2は、当該モジュール１１０-1，１１０-2の中心をなし、クライアント２からの要求を処理する仮想分散ファイルインタフェース１１１-1，１１１-2と、後述するローカルファイルシステム１３-1，１３-2とのインタフェース（ローカルファイルインタフェース）１１２-1，１１２-2と、当該モジュール１１０-1，１１０-2間での通信（後述するマッピングテーブル１４-1，１４-2の情報、及びサーバ情報の一致化のための通信に代表されるサーバ間通信）を行う通信モジュール１１３-1，１１３-2とを持つ。

【００３２】ファイルサーバ１０-1，１０-2はネットワーク３を介してクライアント２と接続されている。ファイルサーバ１０-1，１０-2は、仮想分散ファイルシステム１１の他に、それぞれ当該サーバ１０-1，１０-2に接続されたディスク装置などのストレージ装置１２-1，１２-2（実際のストレージ構成）を管理するローカルなファイルシステム（ローカルファイルシステム）１３-1，１３-2を実装している。

【００３３】ファイルサーバ１０-1，１０-2には、仮想分散ファイルシステム１１とローカルファイルシステム１３-1，１３-2とを対応付ける同一内容のマッピングテーブル１４-1，１４-2が設けられている。このテーブル１４-i（ｉ＝１，２）のデータ構造を図２に示す。

【００３４】テーブル１４-iの各エントリは、仮想分散ファイルシステム１が管理するファイルのファイル名の登録フィールド（ファイル名フィールド）１４１と、当該ファイルの仮想分散ファイルシステム１上の論理的な所在を表す（クライアント２から見える）パス（仮想パ

ス）の登録フィールド（仮想パスフィールド）１４２、当該ファイルのストレージ上の物理的な所在位置を表す（クライアント２から見えない）所在位置情報の登録フィールド（所在位置情報フィールド）１４３、当該ファイルへのアクセス権（許可／禁止）を管理するためのパーミッション情報の登録フィールド（パーミッション情報フィールド）１４４、及びその他の各種属性の登録フィールド１４５を有している。

【００３５】仮想分散ファイルシステム１１（上の仮想分散ファイルモジュール１１０-i）は、このようなデータ構造のマッピングテーブル１４-iを参照することにより、例えばあるファイルがファイルサーバ１０-1，１０-2のいずれにあるか等の所在情報を得ることができる他、パーミッション等、必要に応じてファイルの属性を得ることができる。

【００３６】ファイルサーバ１０-1，１０-2には更に、同一内容のサーバ情報保持部１５-1，１５-2が設けられている。サーバ情報保持部１５-i（ｉ＝１，２）は、図３に示すように、仮想分散ファイルサーバシステム１を構成する全てのファイルサーバ１０-1，１０-2の（ストレージ装置１２-1，１２-2の）空き記憶容量を示す情報（リソース情報）、及び負荷状況を示す情報を含むサーバ情報を保持するのに用いられる。

【００３７】次に図１の構成の動作を説明する。本実施形態では、クライアント２からは、各ファイルサーバ１０-1，１０-2のローカルファイルシステム１３-1，１３-2ではなくて、仮想分散ファイルシステム１１がマウントされているように見えている。そこでクライアント２は、何らかのファイル操作要求が発生した場合、仮想分散ファイルシステム１１が実装されている全ファイルサーバ１０-1，１０-2に対して同一の要求を発行する。この場合、例えばＩＰ（Internet Protocol）マルチキャストを使用する等の手法によれば、クライアント２側はファイルサーバの台数を意識することなく要求の発行が可能である。

【００３８】ファイルサーバ１０-1，１０-2は、クライアント２からの要求を受け取ると、当該要求を仮想分散ファイルシステム１１内の自サーバに対応した仮想分散ファイルモジュール１１０-1，１１０-2に渡す。すると、モジュール１１０-1，１１０-2（内の仮想分散ファイルインタフェース１１１-1，１１１-2）は、その要求がファイルの読み出し要求もしくは書き込み（更新）要求であるか、または新規ファイルの作成要求もしくはディレクトリの作成要求であるか、その要求種別を判別する。

【００３９】ここで、クライアント２からのファイル操作要求がファイルの読み出し要求もしくは書き込み要求であるものとする。この要求には、要求の対象となるファイルのファイル名と、当該ファイルの仮想分散ファイルシステム１１上のパス（仮想パス）が付されている。

【００４０】仮想分散ファイルモジュール１１０-1，１１０-2（内の仮想分散ファイルインタフェース１１１-1，１１１-2）は、クライアント２からのファイル操作要求がファイルの読み出し要求もしくは書き込み要求の場合、要求されたファイルのファイル名及び仮想パスにより自サーバ内のマッピングテーブル１４-1，１４-2を参照し、当該ファイル名及び仮想パスを持つテーブル１４-1，１４-2内エントリ中の所在位置情報フィールド１４３の登録情報から、操作要求のあったファイルが自サーバ内（自サーバに接続されたストレージ装置１２-1，１２-2）に保持されているか否かを調べる。

【００４１】もし、要求されたファイルが自サーバ内に保持されている場合には、仮想分散ファイルモジュール１１０-1，１１０-2（内の仮想分散ファイルインタフェース１１１-1，１１１-2）は、ローカルファイルインタフェース１１２-1，１１２-2により自サーバ内のローカルファイルシステム１３-1，１３-2を介して実際のファイルにアクセスし、クライアント２に応答を返す。一方、操作要求のあったファイルが自サーバ内になかった場合には、他のサーバが応答するものと見なして応答しない。

【００４２】これに対し、クライアント２からの要求が新規ファイルの作成、或いはディレクトリの作成であった場合には、仮想分散ファイルモジュール１１０-1，１１０-2は、（マッピングテーブル１４-1，１４-2ではなくて）サーバ情報保持部１５-1，１５-2を参照する。そして、サーバ情報保持部１５-1，１５-2に保持されている全サーバのサーバ情報をもとに、所定のアルゴリズムに従い、いずれか１つのサーバ１０-i（ｉは１または２）上の仮想分散ファイルモジュール１１０-iだけが、仮想分散ファイルインタフェース１１１-iによりクライアント２からの要求を受け付ける。具体的には、サーバ１０-i上の仮想分散ファイルモジュール１１０-iは、全サーバのサーバ情報の示す空き記憶容量を比較し、自サーバ１０-i（のストレージ装置１２-i）の空き記憶容量が最も大きいと判定できる場合に、クライアント２からの要求を受け付けるものとする。この場合、必ずしもサーバ情報中に負荷状況の情報を持たせる必要はない。

【００４３】なお、全サーバのサーバ情報の示す負荷を比較し、自サーバの負荷が最も低い場合にクライアント２からの要求を受け付けるようにしてもよい。この場合、必ずしもサーバ情報中に対応するサーバの空き記憶容量の情報を持たせる必要はない。

【００４４】この他に、マッピングテーブル１４-iの各ファイル毎のマッピング情報から（つまり各ストレージ装置１２-1，１２-2の領域の使用状況から）、各ストレージ装置１２-1，１２-2上に確保可能な連続領域を求め、必要なサイズ以上の連続領域が確保でき、且つそのサイズが最も大きいストレージ装置が自サーバのストレージ装置１２-iの場合に、クライアント２からの要求を

受け付けるようにしてもよい。この場合、サーバ情報保持部１５-1，１５-2は必ずしも必要でない。

【００４５】更に、空き記憶容量と負荷状況と確保できる連続領域のサイズの少なくとも２つを条件（複合条件）として評価値を求め、自サーバが要求を受け付ける最適なサーバであるか否かを判断ようにしてもよい。

【００４６】さて、ファイルサーバ１０-i上の仮想分散ファイルモジュール１１０-iでは、仮想分散ファイルインタフェース１１１-iによりクライアント２からの要求を受け付けると、要求された新規ファイルの作成、或いはディレクトリの作成を、ローカルファイルインタフェース１１２-iを介してローカルファイルシステム１３-iにより行い、マッピングテーブル１４-1，１４-2に該当するエントリ情報を登録する。

【００４７】新規ファイルの作成、或いはディレクトリの作成が完了した後は、ファイルサーバ１０-i上の仮想分散ファイルモジュール１１０-iでは、自サーバ上のマッピングテーブル１４-iに登録した新たなエントリ情報を通信モジュール１１３-iによりネットワーク３を介して他の全てのサーバ１０-j（ｊは１または２、但しｊ≠ｉ）上の仮想分散ファイルモジュール１１０-ｊに送る。仮想分散ファイルモジュール１１０-ｊ（内の仮想分散ファイルインタフェース１１１-j）は、仮想分散ファイルモジュール１１０-iから送られたマッピングテーブル１４-iのエントリ情報を通信モジュール１１３-jを介して受け取る。そしてモジュール１１０-ｊ（内のインタフェース１１１-j）は、受け取った他サーバ１０-iのマッピングテーブル１４-iのエントリ情報を自サーバ内のマッピングテーブル１４-jに登録する。このように、ファイルサーバ１０-1，１１０-2上の仮想分散ファイルモジュール１１０-1，１１０-2が相互にマッピングテーブル１４-1，１４-2の新規登録されたエントリ情報（更には更新されたエントリ情報）を交換し合うことで、当該マッピングテーブル１４-1，１４-2の内容の一致化を図ることができる。

【００４８】また、ファイルサーバ１０-1，１０-2上の仮想分散ファイルモジュール１１０-1，１１０-2は、自サーバ上のサーバ情報保持部１５-1，１５-2に保持されている各サーバのサーバ情報のうち、自サーバのサーバ情報（空き記憶容量、及び負荷状況）を定期的に更新すると共に、その更新したサーバ情報を（通信モジュール１１３-1，１１３-2により）ネットワーク３を介して他の全てのサーバ（上の仮想分散ファイルモジュール１１０-2，１１０-1）に定期的に送ることで、各ファイルサーバ１０-1，１０-2のサーバ情報保持部１５-1，１５-2の内容の一致化を図る。つまり仮想分散ファイルモジュール１１０-1，１１０-2は定期的にサーバ情報を交換し合うことで一致化を図る。

【００４９】以上の動作によって、ファイルサーバ１０-1，１０-2を自律的に分散・協調動作させることができ、クライアント２には実際にはファイルサーバが２台（複数台）あることを意識させずに、仮想的なファイルサーバを提供することができる。

【００５０】なお、図１のシステムの例ではサーバが２台である場合について説明したが、サーバが３台以上であっても同様の仕組みによって、仮想的なファイルサーバを提供することができる。

【００５１】［第２の実施形態］図４は本発明の第２の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図１と同一部分には同一符号を付してある。

【００５２】図４において、仮想分散ファイルサーバシステム４の中心をなす仮想分散ファイルシステム４１は、ｎ台のファイルサーバ１０-1～１０-nに分散して実装されている。この仮想分散ファイルシステム４１は、図１中の仮想分散ファイルシステム１１と同様に、全ファイルサーバ１０-1～１０-nのファイルを統合的に管理し、各ファイルサーバ１０-1～１０-nそれぞれの実際のボリューム構成には依存しない、仮想的なファイルシステムをクライアント２に対して提供している。仮想分散ファイルシステム４１は、各ファイルサーバ１０-1～１０-n上でクライアント２からの要求を処理する仮想分散ファイルモジュール４１０-1～４１０-nを有している。モジュール４１０-1～４１０-nは、図１中のモジュール１１０-1，１１０-2と同様の構成である、仮想分散ファイルインタフェース４１１-1～４１１-nと、ローカルファイルインタフェース４１２-1～４１２-nと、通信モジュール４１３-1～４１３-nとを持つ。但し、本実施形態における通信モジュール４１３-1～４１３-nは、図１中の通信モジュール１１３-1，１１３-2と異なり、後述するプライベート通信路５を介して通信を行うように構成されている。

【００５３】ファイルサーバ１０-1～１０-nはネットワーク３を介してクライアント２と接続されている。ファイルサーバ１０-1～１０-nは、仮想分散ファイルシステム１１の他に、それぞれ当該サーバ１０-1～１０-nに接続されたストレージ装置１２-1～１２-2を管理するローカルなファイルシステム（ローカルファイルシステム）１３-1～１３-2を実装している。ファイルサーバ１０-1～１０-nには、マッピングテーブル１４-1～１４-nと、サーバ情報保持部１５-1～１５-2とが設けられている。

【００５４】図４の構成の仮想分散ファイルサーバシステム４の特徴は、図１の構成の仮想分散ファイルサーバシステム１と異なって、システムを構成するファイルサーバの台数がｎ台である点と、そのｎ台のファイルサーバ１０-1～１０-nがネットワーク３とは別のプライベートな通信路５によっても相互接続されている点である。このプライベート通信路５は、例えばイーサネット、或いはファイバチャネル（Fibre Channnel）等であるが、

物理層に関しては特定しない。またトポロジに関しても、図４の例ではバス型を想定しているが、ループやスイッチであってもよい。

【００５５】図４の構成において、ファイルサーバ１０-1～１０-nが（仮想分散ファイルシステム４１内の仮想分散ファイルモジュール４１０-1～４１０-nにより）分散・協調動作を行うためには、前記第１の実施形態での動作説明から類推されるように、マッピングテーブル１４-1～１４-n、及びサーバ情報保持部１５-1～１５-nの内容を、各サーバ１０-1～１０-n間で常に一致化させておく必要がある。しかし、サーバ１０-1～１０-n間の情報一致化を、前記第１の実施形態と同様にネットワーク３を介して行うのでは、仮想分散ファイルサーバシステム４を構成するファイルサーバの台数が増加した場合には、その情報一致化の（ためのサーバ間通信の）トラフィックが増加し、ネットワーク３上のスループットを悪化させることになる。

【００５６】そこで本実施形態（第２の実施形態）では、図４の構成のように、各ファイルサーバ１０-1～１０-n間にサーバ間の情報交換専用のプライベート通信路５を設け、仮想分散ファイルシステム４１内の仮想分散ファイルモジュール４１０-1～４１０-nで通信モジュール４１３-1～４１３-nにより行われるサーバ間通信に、即ちマッピングテーブル１４-1～１４-n、及びサーバ情報保持部１５-1～１５-nの内容を一致化するためのサーバ間通信に、この通信路５を使用するようにしている。

【００５７】このように本実施形態では、マッピングテーブル１４-1～１４-n、及びサーバ情報保持部１５-1～１５-nの内容の一致化のためのサーバ間通信に、ネットワーク３でなくてプライベート通信路５を用いることにより、ネットワーク３の負荷の軽減を図ることができる。

【００５８】［第３の実施形態］以上に述べた第１、第２の実施形態では、複数のファイルサーバを分散・協調動作させる仮想分散ファイルサーバシステムの構成例を示した。この第１、第２の実施形態で参照した図１、図４の構成は、特定のサーバ台数における静的な例である。しかし、サーバ台数については、変更可能な構成とすることが好ましい。

【００５９】そこで、仮想分散ファイルサーバシステムを構成するサーバ台数を動的に拡張可能とした本発明の第３の実施形態について図面を参照して説明する。図５は本発明の第３の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図４と同一部分には同一符号を付してある。

【００６０】まず、図４に示した仮想分散ファイルサーバシステム４、即ちｎ台のファイルサーバ１０-1～１０-nで構成される仮想分散ファイルサーバシステム４に、図５（ａ）に示すように、新たなファイルサーバ１０-(n+1)を追加するものとする。

【００６１】この場合、追加されたファイルサーバ１０-(n+1)にも分散されている仮想分散ファイルシステム４１上の仮想分散ファイルモジュール４１０-(n+1)は、既に仮想分散ファイルサーバシステム４を構成しているファイルサーバ１０-1～１０-nに対し、図５（ａ）において符号Ａ１で示すように、（例えば図示せぬプライベート通信路を介しての）サーバ間通信により、マッピングテーブル１４-1～１４-nのエントリ情報及びサーバ情報保持部１５-1～１５-nのサーバ情報（各サーバのリソース情報及び負荷状況を含む）の更新をロックする。

【００６２】その上で、追加されたサーバ１０-(n+1)上のモジュール４１０-(n+1)は、他のファイルサーバ１０-1～１０-nのうちのいずれかのサーバ、例えばファイルサーバ１０-1から、図５（ｂ）において符号Ａ２で示すように、マッピングテーブル１４-1及びサーバ情報保持部１５-1の全情報を、サーバ間通信により自サーバ内のマッピングテーブル１４-(n+1)及びサーバ情報保持部１５-(n+1)にコピーする。

【００６３】次に、追加されたファイルサーバ１０-(n+1)上のモジュール４１０-(n+1)は、コピー後のサーバ情報保持部１５-(n+1)に対し、図５（ｃ）において符号Ａ３で示すように、自サーバのリソース及び負荷状況を示すサーバ情報を追加する。

【００６４】しかる後にファイルサーバ１０-(n+1)上のモジュール４１０-(n+1)は、図５（ｄ）において符号Ａ４で示すように、サーバ間通信により他の全ファイルサーバ１０-1～１０-nに対してサーバ情報の一致化要求を発行し、その後にロックを解除する。

【００６５】以上の一連の動作により、既に構築されている仮想分散ファイルサーバシステム４に対して、動的に薪たなサーバ（ファイルサーバ１０-(n+1)）を追加することができる。この場合、例えば現在の仮想分散ファイルサーバシステム４のボリューム構成に対し、新規リソースをどのように振り分けるか、といった情報を付加すれば、必要に応じてボリュームを選択的に拡張することも可能である。

【００６６】［第４の実施形態］図６は本発明の第４の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図４と同一部分には同一符号を付してある。

【００６７】図６において、６は図４中の仮想分散ファイルサーバシステム４に相当する仮想分散ファイルサーバシステムである。この仮想分散ファイルサーバシステム６の特徴は、当該システム６を構成するファイルサーバ１０-1～１０-n内に、自サーバ（のストレージ装置１２-1～１２-n）に保持されている各ファイルについての負荷状況の情報（ファイル別負荷状況情報）を保持するファイル別負荷状況情報保持部１６-1～１６-nを備えて

いる点にある。これに伴い、仮想分散ファイルサーバシステム６の中心をなす仮想分散ファイルシステム６１の持つ機能も図４中の仮想分散ファイルシステム４１とは一部異なる。但し、仮想分散ファイルシステム６１内の各ファイルサーバ１０-1～１０-n毎の仮想分散ファイルモジュールには便宜的に図４と同一符号（４１０-1～４１０-n）を用いている。なお、図６では、モジュール４１０-1～４１０-n内の構成要素（仮想分散ファイルインタフェース、ローカルファイルインタフェース、通信モジュール）、及びファイルサーバ１０-1～１０-n上のマッピングテーブル、サーバ情報保持部は省略されている。

【００６８】ファイル別負荷状況情報保持部１６-i（i＝1～n）は、図７（a）に示すデータ構造を持ち、自サーバ内のファイル毎の負荷状況を示す情報と、そのファイルの属性を示す情報（ファイル属性）と、レプリケーションフラグとを含むファイル別負荷状況情報を保持する。ファイル属性は、対応するファイルがオリジナルであるかレプリカ（複製）であるかを示す。また、レプリケーションフラグは、対応するフラグがオリジナルの場合、そのレプリカを他サーバ側に生成済みであるか否か、つまりレプリケーション済みであるか否かを示す。

【００６９】図６には、ファイルサーバ１０-nが持つストレージ装置１２-n内に、ファイルサーバ１０-1が持つストレージ装置１２-1内の任意のファイル６１１のレプリカ６１２がレプリケーションＢ１により保持されている様子が示されている。

【００７０】次に、図６の構成の動作を説明する。仮想分散ファイルシステム６１上の各仮想分散ファイルモジュール４１０-1～４１０-nは、ファイル別負荷状況情報保持部１６-1～１６-nを例えば定期的に参照する。そしてモジュール４１０-1～４１０-nは、保持部１６-1～１６-nに保持されているファイル別の負荷状況情報から、自サーバ（のストレージ装置１２-1～１２-n）に保持されているファイルの中に、第１の閾値を超えた負荷のファイルが存在することを検出した場合、他サーバの１つに対して対応するファイルのレプリカを非同期に生成するレプリケーション動作を、例えばプライベート通信路５を介してのサーバ間通信により行う。ここでファイルの負荷状況は、当該ファイルへの要求の待ち行列（キュー）にある要求数、或いは当該ファイルの待ち状態にある要求の示すサイズの総和であり、要求を受け付ける毎と要求を処理し終える毎に更新される。また、レプリケーションの対象サーバには、図示せぬサーバ情報保持部に保持されているサーバ情報に基づいて、例えば負荷が最も低いサーバを選択すればよい。

【００７１】仮想分散ファイルモジュール４１０-1～４１０-nはレプリケーション動作を行うと、自サーバのファイル別負荷状況情報保持部１６-1～１６-nに保持されている対応するファイルの負荷状況情報中のレプリケーションフラグをレプリケーション済みの通知状態にセットする。またレプリケーション動作の対象となったサーバの仮想分散ファイルモジュールは、自サーバのファイル別負荷状況情報保持部内に対応するレプリカの負荷状況情報を追加する。

【００７２】ここでは、図６に示すように、ファイルサーバ１０-1の保持するファイル６１１のレプリケーションＢ１がプライベート通信路５を介してファイルサーバ１０-nに対して行われて、そのレプリカ６１２が当該ファイルサーバ１０-nのストレージ装置１２-nに保持されたものとする。この場合、ファイルサーバ１０-1のファイル別負荷状況情報保持部１６-1に保持されているファイル６１１の負荷状況情報中のレプリケーションフラグがレプリケーション済みを示す状態にセットされる。また、ファイルサーバ１０-nのファイル別負荷状況情報保持部１６-nには、ファイル６１１のレプリカ６１２についての新たな負荷状況情報が追加される。この負荷状況情報中のファイル属性は、対応するファイルが（ファイル６１１の）レプリカ（６１２）であることを示す。

【００７３】以後、クライアント２からファイル６１１の新たな読み出し要求があった場合、当該ファイル６１１を保持するファイルサーバ１０-1（の仮想分散ファイルモジュール４１０-1）は、当該ファイル６１１の負荷が第２の閾値（但し、第２の閾値＜第１の閾値）を超えているか否かを調べ、超えているならば、クライアント２からの要求に応答しない。この場合、クライアント２からの要求に対しては、レプリケーションを受けたファイルサーバ１０-nが応答する。ここでファイルサーバ１０-nは、ファイルサーバ１０-1が応答するか否かを考慮する必要はなく、要求されたファイル６１１のレプリカ６１２を有する限り、クライアント２に応答すればよい。

【００７４】このように、クライアント２からのファイル６１１に対する新たな読み出し要求を、そのレプリカ６１２を用いてファイルサーバ１０-nが処理することで、そのファイル６１１を保持するファイルサーバ１０-1では、それ以前に受け付けた当該ファイル６１１に対する読み出し要求の処理が進み、当該ファイル６１１の負荷が上記第２の閾値以下となる。するとファイルサーバ１０-1上の仮想分散ファイルモジュール４１０-1は、ファイルサーバ１０-n上の仮想分散ファイルモジュール４１０-nに対して、ファイル６１１のレプリカ６１２を消去するための要求を例えばプライベート通信路５を介したサーバ間通信により送る。

【００７５】この要求を受けたファイルサーバ１０-n上の仮想分散ファイルモジュール４１０-nは、既に受け付け済みの要求に対してのみレプリカ６１２を用いて処理を行い、しかる後にレプリカ６１２と対応する負荷状況情報を消去する。一方、ファイルサーバ１０-1上の仮想分散ファイルモジュール４１０-1は、クライアント２か

らのファイル６１１に対する新たな読み出し要求があれば、それに対して応答する。

【００７６】ところで、ファイルサーバ１０-1からファイルサーバ１０-nへのファイル６１１のレプリケーションにより、当該ファイル６１１に対する読み出し要求をファイルサーバ１０-nで受け付けるようになった結果、ファイルサーバ１０-1におけるファイル６１１の負荷が第２の閾値以下となる前に、ファイルサーバ１０-nにおける当該ファイル６１１のレプリカ６１２の負荷が第１の閾値を超えることがあり得る。

【００７７】そこで、このような場合、今度はファイルサーバ１０-nがレプリカ６１２を用いて次の世代のレプリカを他の１つのサーバに生成し、即ちレプリケーションのレプリケーションを行い、そのサーバでファイル６１１に対する読み出し要求を処理させればよい。そのためには、ファイル別負荷状況情報保持部１６-i（ｉ＝１～ｎ）に保持されるファイル毎の負荷状況情報に、図７（ａ）に示したような負荷状況とレプリケーションフラグに加えて、図７（ｂ）に示すように、ファイルの世代情報を持たせるとよい。

【００７８】この場合、ある世代のレプリカの負荷が上記第２の閾値以下に下がった時点で、当該レプリカを持つサーバから、そのサーバによるレプリケーションの対象となったサーバの持つ次世代のレプリカを消去する等の制御を行うことができる。このとき、レプリカの消去が要求されたサーバが、別のサーバに対して更に次世代のレプリカを生成している場合、その更に次世代のレプリカを消去するとよい。この他に、少なくともレプリカに関連したファイルの負荷状況情報については、前記したサーバ情報と同様に、各サーバ間の一致化を図ることにより、同一ファイル（レプリカを含む）について、負荷が最も低いファイルを持つサーバが、当該ファイルに対する読み出し要求に応答するようにしてもよい。

【００７９】最近では、ビデオ、オーディオ等のストリーミングデータや、或いはWWW（World Wide Web）のコンテンツ等、基本的には読み出しが主で、比較的サイズが大きく、ある程度のレスポンス（場合によっては帯域保証）が必要なデータが増加しつつある。しかもこうしたデータは、短期的に見て特定のデータ（ファイル）にアクセスが集中するケースが想定されるため、レスポンスを確保するのが困難な場合もある。以上に述べた図６の構成は、こうした状況を想定したもので、特定のファイルにアクセスが集中した場合に、自動的に当該ファイルのレプリケーションを行うことで、当該ファイルへのアクセスを分散させることができるようにしている。この構成は、単に負荷分散だけでなく、例えば重要性の高いファイルのバックアップに利用することも可能である。

【００８０】［第５の実施形態］図８は本発明の第５の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図であり、図４と同一部分には同一符号を付してある。

【００８１】図８において、８は図４中の仮想分散ファイルサーバシステム４に相当する仮想分散ファイルサーバシステムである。この仮想分散ファイルサーバシステム８の特徴は、ファイルサーバ１０-1～１０-n、及びストレージ装置１２-1～１２-nが、例えばＦＣ－ＡＬ（Fibre Channnel Arbitrated Loop）８０により相互接続され、（ホストとしての）各ファイルサーバ１０-1～１０-nから（ターゲットとしての）ストレージ装置１２-1～１２-nの共有が可能な（つまりマルチホスト可能な）ネットワーク構成を適用している点にある。ここでは、図４の構成と異なって、プライベート通信路５を持たない点に注意されたい。

【００８２】この図８の構成では、図４の構成において（仮想分散ファイルモジュール４１０-1～４１０-nの通信モジュール４１３-1～４１３-nにより）プライベート通信路５を介して行われるサーバ間通信を、図１の構成と同様にネットワーク３を介して行えばよい（図は、この状態が示されている）。また、上記サーバ間通信を、ファイルサーバ１０-1～１０-nのストレージ接続用のインタフェースを介してＦＣ－ＡＬ８０上で行うようにしてもよい。この場合、プライベート通信路５を用いたのと同様に、ネットワーク３の負荷を軽減できる。

【００８３】図８の構成によれば、ストレージ装置１２-1～１２-nが全てのファイルサーバ１０-1～１０-nから直接に見えるので、各サーバ１０-1～１０-nに図６中のファイル別負荷状況情報保持部１６-1～１６-nを持たせることで、前記第４の実施形態で述べたようなレプリケーション動作や負荷分散を容易に行うことができる。なお、マルチホスト可能なネットワーク（インタフェース）はＦＣ－ＡＬ８０に限るものではなく、ＳＣＳＩ（Small Computer System Interface）バスであっても構わない。

【００８４】

【発明の効果】以上詳述したように本発明によれば、ネットワーク上に分散した複数のファイルサーバを、クライアントからは単一のサーバとして扱うことができ、サーバ台数やストレージ装置の接続状態をクライアントに意識させることがない。

【００８５】また本発明によれば、サーバを増設した場合、動的にボリュームを拡張することもできる。

【００８６】更に本発明によれば、複数のサーバ間で自律的な負荷分散が実現できる。

【図面の簡単な説明】

【図１】本発明の第１の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図２】図１中のマッピングテーブルのデータ構造例を

示す図。

【図３】図１中のサーバ情報保持部のデータ構造例を示す図。

【図４】本発明の第２の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図５】本発明の第３の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図６】本発明の第４の実施形態に係る仮想分散ファイルサーバシステムを適用するコンピュータ・ネットワークシステムの構成を示すブロック図。

【図７】図６中のファイル別負荷状況情報保持部のデータ構造例を示す図。

【図８】同実施形態の動作を説明するタイミングチャート。

【符号の説明】

１，４，６，８…仮想分散ファイルサーバシステム

２…クライアント

３…ネットワーク

５…プライベート通信路

１０-1～１０-n…ファイルサーバ

１１，４１，６１…仮想分散ファイルシステム

１２-1～１２-n…ストレージ装置

１３-1～１３-n…ローカルファイルシステム

１４-1～１４-n…マッピングテーブル

１５-1～１５-n…サーバ情報保持部

１６-1～６-n…ファイル別負荷状況情報保持部

８０…ＦＣ－ＡＬ（マルチホスト可能なインタフェース）

１１０-1～１１０-n，４１０-1～４１０-n…仮想分散ファイルモジュール（管理モジュール）

１１１-1～１１１-n…仮想分散ファイルインタフェース

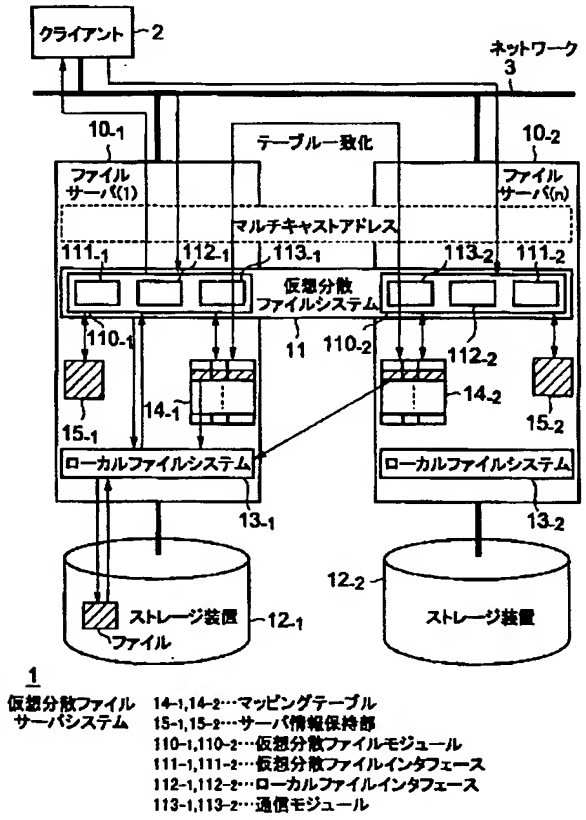
１１２-1～１１２-n…ローカルファイルインタフェース

１１３-1～１１３-n…通信モジュール

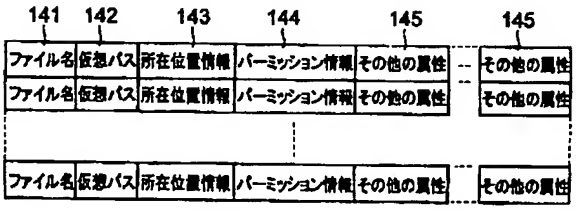
６１１…ファイル

６１２…（ファイル６１１の）レプリカ

【図１】



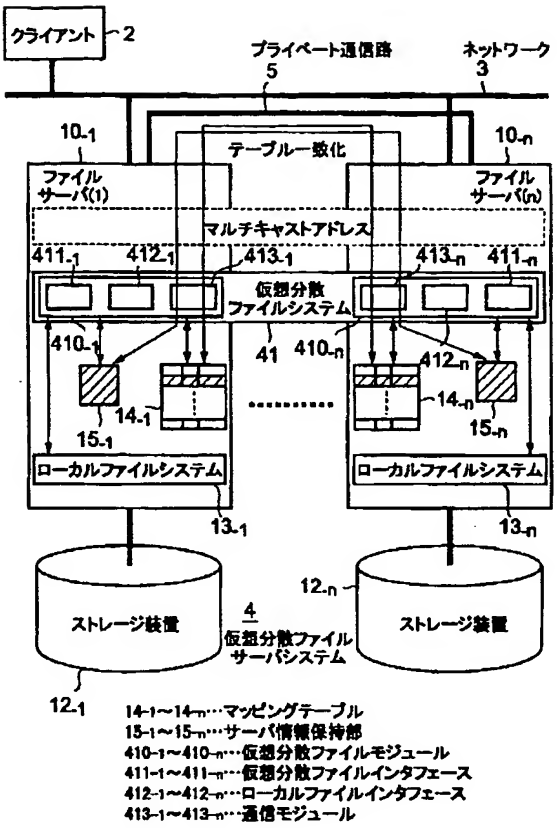
【図２】



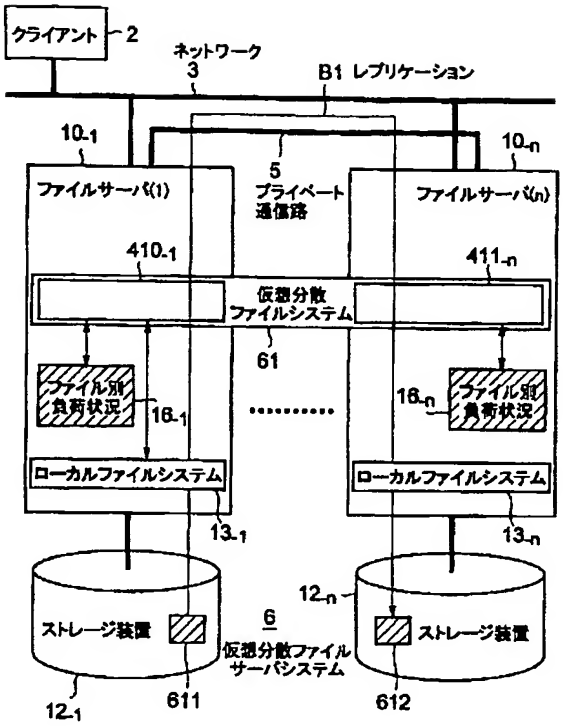
【図３】

サーバID	リソース情報	負荷状況
サーバ1		
サーバ2		

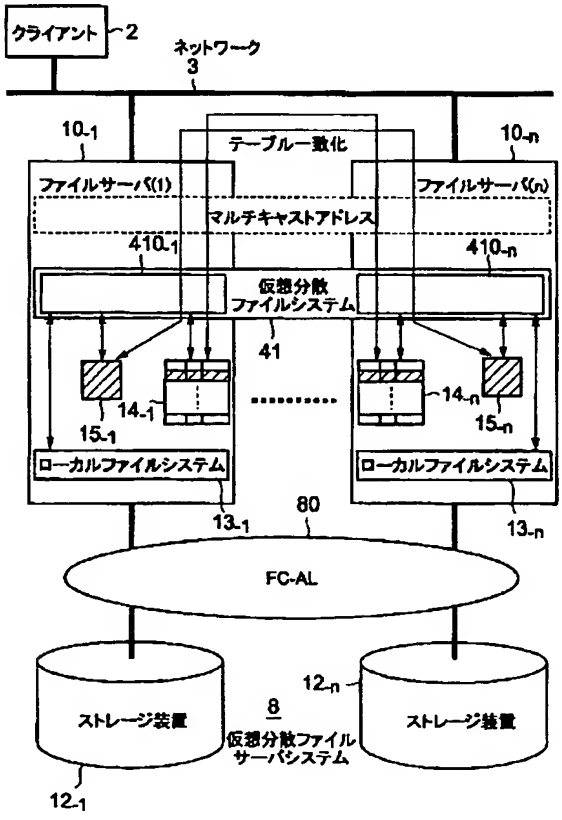
【図４】


クライアント 2
プライベート通信路 5
ネットワーク 3
10-1
10-n
テーブル一致化
ファイルサーバ(1)
ファイルサーバ(n)
マルチキャストアドレス
411-1
412-1
413-1
413-n
411-n
仮想分散ファイルシステム
41
410-1
410-n
412-n
14-1
14-n
15-1
15-n
ローカルファイルシステム
13-1
13-n
12-n
ストレージ装置
4
仮想分散ファイルサーバシステム
12-1
14-1～14-n…マッピングテーブル
15-1～15-n…サーバ情報保持部
410-1～410-n…仮想分散ファイルモジュール
411-1～411-n…仮想分散ファイルインタフェース
412-1～412-n…ローカルファイルインタフェース
413-1～413-n…通信モジュール

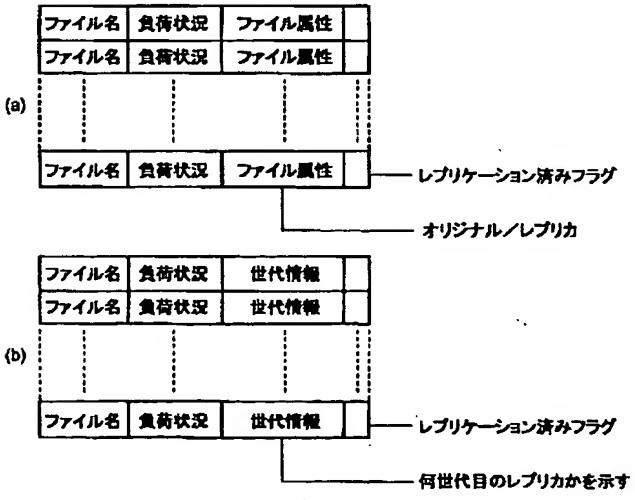
【図６】


クライアント 2
ネットワーク 3
B1 レプリケーション
10-1
10-n
ファイルサーバ(1)
5
プライベート通信路
ファイルサーバ(n)
410-1
411-n
仮想分散ファイルシステム
61
ファイル別負荷状況
16-1
16-n
ローカルファイルシステム
13-1
13-n
12-n
ストレージ装置
6
仮想分散ファイルサーバシステム
12-1
611
612

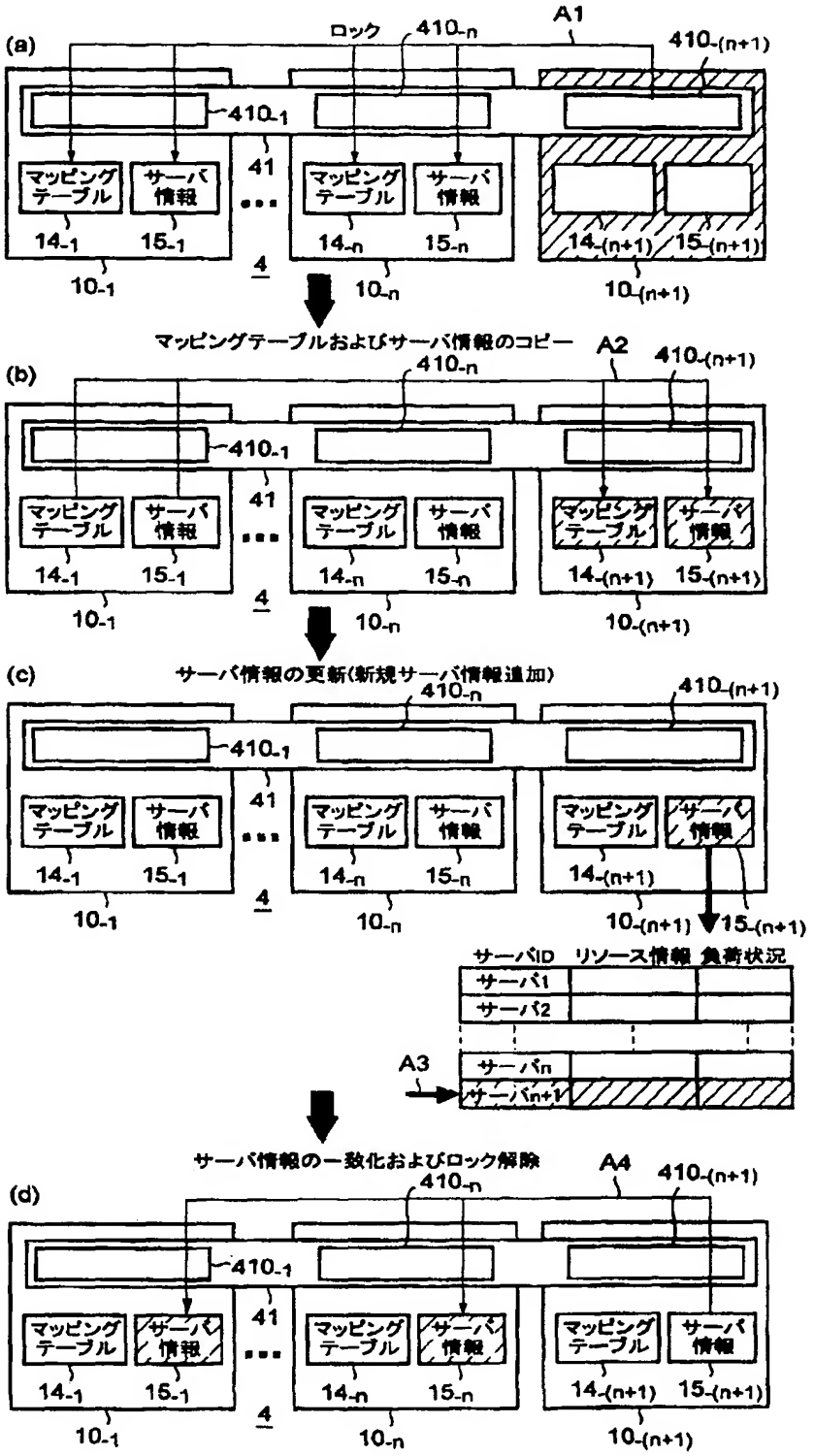
【図８】


クライアント 2
ネットワーク 3
10-1
10-n
テーブル一致化
ファイルサーバ(1)
ファイルサーバ(n)
マルチキャストアドレス
410-1
410-n
仮想分散ファイルシステム
41
14-1
14-n
15-1
15-n
ローカルファイルシステム
13-1
80
13-n
FC-AL
12-n
ストレージ装置
8
仮想分散ファイルサーバシステム
12-1

【図７】


(a)
ファイル名
負荷状況
ファイル属性
レプリケーション済みフラグ
オリジナル／レプリカ
(b)
世代情報
何世代目のレプリカかを示す

【図５】


(a)
ロック
410_n
A1
410_(n+1)
410_1
マッピングテーブル
サーバ情報
41
14_1
15_1
14_n
15_n
14_(n+1)
15_(n+1)
10_1
4
10_n
10_(n+1)
マッピングテーブルおよびサーバ情報のコピー
(b)
A2
(c)
サーバ情報の更新(新規サーバ情報追加)
サーバID
リソース情報
負荷状況
サーバ1
サーバ2
サーバn
サーバn+1
A3
サーバ情報の一致化およびロック解除
(d)
A4

フロントページの続き

Ｆターム(参考) 5B082 CA18 EA07 HA03 HA05 HA08
HA09
5B089 GA12 JA11 JB15 KA00 KC15
KC28 KE07